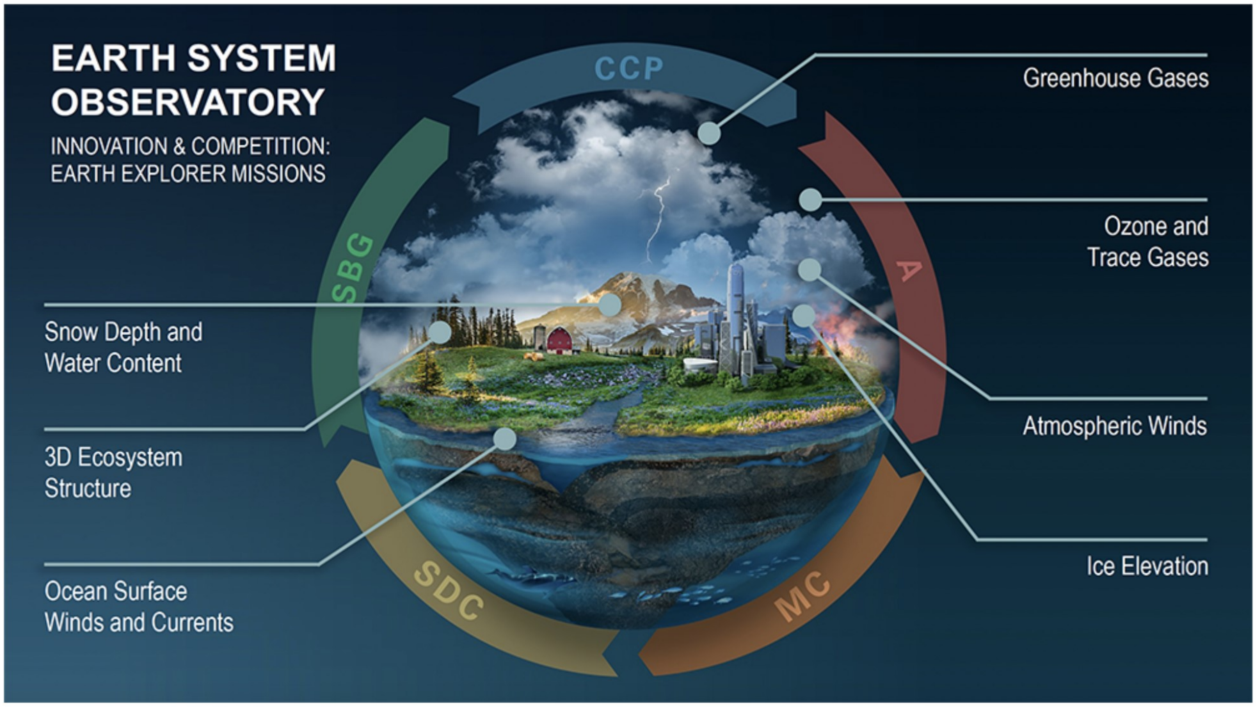
Workshop #1 Report

ESO Mission Data Processing Study - Summary of NASA Program Offices and ESO Missions Requirements, Constraints, Recommendations, and Opportunities



[A portion of] this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004).

Contributors

E. Natasha Stavros (Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder)

Elias Sayfi (Jet Propulsion Laboratory (JPL), California Institute of Technology)

Bernie Bienstock (Jet Propulsion Laboratory (JPL), California Institute of Technology)

Wenying Su (NASA Langley Research Center)

Hook Hua (Jet Propulsion Laboratory (JPL), California Institute of Technology)

Andrew Michaelis (NASA Ames Research Center)

Evelyn Ho (NASA Goddard Space Flight Center)

Karen Yuen (Jet Propulsion Laboratory (JPL), California Institute of Technology)

Qing Yue (Jet Propulsion Laboratory (JPL), California Institute of Technology)

Curt Tilmes (NASA Goddard Space Flight Center)

Lesley Ott (NASA Goddard Space Flight Center)

Chris Engebretson (USGS)

Adrian Parker (NOAA/NESDIS)

Sean Harkins (Marshall Space Flight Center)

Mike Chepurin (NOAA/NESDIS)

Table of Contents

Purpose and Scope

The ***purpose*** of this report is to synthesize findings from the NASA Earth System Observatory (ESO; Margetta, 2021) Processing Workshop #1. The full agenda and announcement is found here: https://earthdata.nasa.gov/esds/open-science/oss-for-eso-workshops.

This workshop is the first of three in the “Open Source Science For ESO Mission Data Processing Study”. Workshop #1 focused on gathering needs and considerations for evaluating different open science data system architectures to support Earth system science and mission data system efficiencies. The goals of workshop #1 were to:

1. understand programmatic requirements, objectives, capability needs, and constraints driving ESO Mission Data Processing Systems (MDPS; see glossary);
2. seek opportunities to advance the science data systems in the context of the Open Source Science for ESO Mission Data Processing Study objectives; and
3. establish programmatic and mission point of contacts in support of codevelopment of future ideas and concepts.

This document acts as a documentation of key points made during the workshop by study stakeholders, and an open and transparent mechanism for clearly communicating definition of evaluation criteria as used by the System Architecture Working Group (SAWG) for evaluating different architectures in later phases of the architecture study.

The ***scope*** of this report focuses on synthesizing the information provided by workshop #1 participants and documenting a path forward for the SAWG by contextualizing these inputs in the larger study.

Reference Documents and Materials

Study Website with links to workshop agendas, presentations, and related documents: https://earthdata.nasa.gov/esds/open-science/oss-for-eso-workshops

Transform to Open Science Github: https://github.com/nasa/Transform-to-Open-Science

Executive Summary

The purpose of the ESO Mission Data Processing Study is to identify the architectures that best:

1. Meet the ESO mission science data processing objectives
2. Enable data system efficiencies
3. Support Earth system science and applications
4. Promote open science principles to expand participation in mission science beyond the funded science teams

Sponsored by Kevin Murphy, Chief Science Data Officer, NASA Science Mission Directorate (SMD), and Program Manager, Earth Science Division (ESD), it's focused on the four ESO missions: NISAR/SDC, SBG, MC, and AOS. The study team consists of the Steering committee whose primary role is management of the study and the System Architecture Working Group (SAWG) responsible for conducting the Mission Data Processing System (MDPS) architecture trade analysis. The study is conducted via a series of three workshops. Workshop #1 focused on understanding the NASA Program goals and ESO mission needs for the purpose of evaluating different data system architectures. This document is a report of the findings of Workshop #1.

Stakeholder Priorities. From the NASA programmatic perspective, the ESD priorities are to advance scientific discovery through open source science (OSS) and foster broad participation by NASA data users earlier in the mission scientific development process, starting in Phase B. The Flight Project Program strives to meet those objectives while still adhering to data system design constraints and flight requirements, keeping compliance with existing concepts and investments, and enforcing the role of DAACs as “Science Enabling Centers”. The R&A Program sees the accommodation of the breadth of activities across many data sources as crucial in supporting the larger researcher community, and enabling community-based, intra-team collaboration. The Applied Science Program would like to expand access to mission data and services, to enable co-development approaches, across disciplines, with the broader community of variable skill. The ESTO program supports on-prem (High-End Computing - HEC) and cloud capabilities, Findabale, Accessible, Interoperable, and Reproducible (FAIR) principles for software and data, and the sponsoring of technology advancements for future Earth Science division needs.

The ESO mission program offices see partnering with external and international communities, access to many data sources specially the Program of Record, and support for Analysis-Ready Data as crucial elements. NISAR and SBG are very interested in lowering the bar for data product usage while MC acknowledges that some algorithms are proprietary and not worth sharing.

The ESO missions science teams would benefit from the sharing of data and algorithms, an analysis platform (specially for L3+), and access to POR, hence harmonization. NISAR specifically expressed that open access to data, with quality metadata, is most conducive to scientific progress while AOS planned to use the analysis platform pre-launch with simulated/airborne data.

The ESO MDPS have a need for a community supported on-demand processing system to generate higher level (L3+) products (ARD) that enable harmonization. Data volumes range from small to large (MC at <1PB to NISAR/SBG w/ 100+ PB), while variable processing and low latency products, and interfacing with external MDPSs is essential. All but MC are planning to be co-located with the DAAC in AWS and develop all software in the open.

Stakeholder Considerations and Constraints. Findings relevant to all stakeholders are captured in a Policy, Economics, Sociocultural Factors, and Technologies (PEST; Aguilar, 1967, Stavros, 2021) analysis. Cybersecurity, Intellectual property (IP), ITAR, and SMD policies on Data Management and Open Source Science don't always align, which can be major hurdles to OSS. The data system economics is cost-constrained by ESO budgets, hindered by some in the community who are convinced it's cheaper to buy their own systems, and limited by NASA's investment in HPC. Sociocultural factors include the science community core business that is based on reward for being the first to discover which disincentivizes participation in OSS. A cultural shift is needed to teach experienced researchers new practices, provide cloud/HPC training, and develop consistent OSS guidelines across different centers. Technological considerations include use of existing investments in infrastructure and mechanisms for community contributions, while limiting the proliferation of unvalidated data.

Additionally, a Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis of data systems was done. Strengths of existing data systems are that they make NASA data accessible, support large data and compute infrastructure, and have known costs and schedules. Weaknesses include limitations on community access, limitations on use of existing infrastructure, challenges with interoperability across compute infrastructure, lack of expertise to work with NASA data, and lack of recognition and merit for support of OSS software and tooling. Data system opportunities include capacity building, early adopters, standardization of the value of OSS, use of existing infrastructure, and developing open source software and data tools. Threats to providing an OSS data system include a disconnect between OSS policy and implementation within the existing business model and cybersecurity policies; the potential to introduce inequity in terms of access, inclusion, and work effort; cost-optimization in the cloud and across diverse geographic regions; and research community hesitancy on cloud compute adoption.

Synthesis of Common Themes by Data System Objective. Along with the analysis described above, Common Themes were synthesized from the stakeholder responses and aligned along the four objectives of the study: 1) ESO MPDS are on the cloud and on-prem; support forward, on-demand, and low latency processing; interface with external systems, and are cost-constrained. 2) Efficiency opportunities include DAAC co-location; flexibility/scalability to adapt to varying data volumes/compute needs; and common data formats. 3) Earth system science is advanced by sharing of data/algorithms, supporting multidisciplinary research, and a common architecture that enables cross-ESO science objectives. 4) Promoting open science needs a publicly accessible, extensible analysis platform with access controls that can track metrics both for cost accounting and adoption encouragement. See Table 1 for full details.

The following Evaluation Criteria were derived from the Common Themes, again, aligned with the four objectives of the study: 1) the Data System shall support mission needs, be portable, have well defined interfaces, be relatively mature before use, and be able to be developed within existing budgets; 2) the Data System shall support a data lake, be flexible and efficient, accommodate varying compute needs, and encourage standard data formats; 3) the Data System shall enable data/algorithm/tools sharing to facilitate the advancement of cross-ESO science goals; and 4) the Data System shall provide a community-based, publicly accessible Analysis Platform that is cybersecurity compliant. See Table 2 for full details.

Path Forward. With Workshop #1 complete, and the findings synthesized into this report, Workshop #2 in March 2022 will solicit input from the community of data systems to identify architectures to be evaluated against the Evaluation Criteria, which were derived from the Common Themes, based on their feasibility, maturity, and dependencies. The final results of the study will be presented in Workshop #3.

Highlights

1. An MDPS architecture must support mission requirements and NASA Earth system science and applications.
2. An MDPS architecture should be portable (cloud, on-prem, hybrid) and scalable to support small and large deployments (<5 to 1000+ nodes).
3. Missions must continue to produce quality products and support how those products can integrate with other missions' products.
4. Co-location of data into a datalake is foundational to enabling Earth system science and applications.
5. There needs to be an on-demand analysis platform that democratizes access, collaboration, and use of the MDPS architecture by the broader community.
6. Serving the open science community influences MDPS design, policies, and procedures for researcher engagement.
7. Existing policies (Cybersecurity, ITAR, IP) implementations have limited our ability to serve open science and policy should be evolved.
8. Open-source software fundamentally changed how software was shared and accelerated its impact; we are applying this transformation to all of science (data, software, documentation and knowledge) to expand participation in science and accelerate discovery.
9. Foster Open Source Science by redesigning reward systems, establishing early adopter programs, and capacity building.
10. Open Source Science Principles have to be a core activity.

ESO Study Overview

The motivation for a data system architecture study stems from recognition that access to algorithms, workflows, computing, and analytics has been a major barrier to participating in NASA science. Opening the access provides greater opportunities for more people to participate in NASA science.

The purpose of this study is to assess methods to enable data system efficiency to support the next decade of NASA Earth System Observatory (ESO; Margetta, 2021) that support Earth system science and promote open science principles to expand participation in mission science beyond funded science teams. The focus is not specifically on data archiving, but on how the broader community can participate in mission data system processing (see glossary).

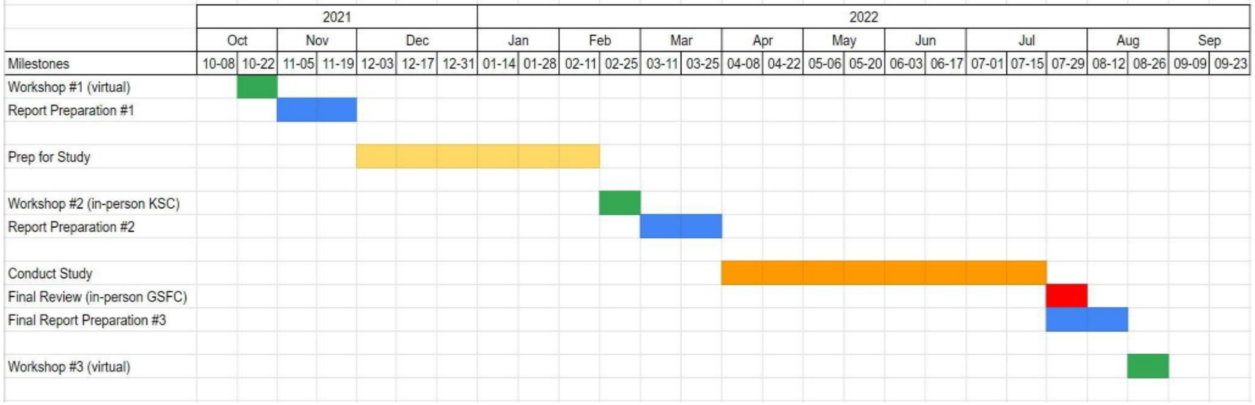
The objectives of the data system architecture study are to identify and assess potential data system architectures that can:

5. Meet the ESO mission science data processing objectives
6. Enable data system efficiencies
7. Support Earth system science and applications
8. Promote open science principles to expand participation in mission science beyond the funded science teams

The principles of this study are to practice open, team science by conducting meetings in the open, thoroughly recording the conversations during workshops, and by making workshop artifacts citable with DOIs and accessible through the Study website and Github. In addition, participants in one-on-one meetings (between the study team and other entities) should document and make notes from the meetings accessible, ensuring community participation, provide mechanisms for continuous feedback, and actively seek feedback from historically excluded communities.

This study is sponsored by Kevin Murphy, Chief Science Data Officer, Science Mission Directorate, and Program Manager, Earth Science Data Systems Program. The study consists of core staff who help gather inputs from a broader community including the Steering Committee and the SAWG. Descriptions of the roles of these core staff are outlined in the glossary. The SAWG is responsible for collecting and evaluating data system architecture drivers including: 1) ESO program goals, constraints, and opportunities; 2) ESO mission objectives and capability needs; 3) the state of the practice in open-science and data processing systems; and 4) community recommendations. The SAWG will perform a trade study that establishes viable architectural options and implementation approaches. To accomplish this, they will establish evaluation criteria (qualitative and quantitative) for use in the analysis of the trade space. The SAWG will use the trade study to provide candidate architectures and make recommendations. All methods and findings will be clearly documented.

The approach of the study is to solicit stakeholder feedback through open workshops and public Requests for Information (RFI). There are three workshops planned. Workshop #1 focused on understanding the NASA program goals and ESO mission needs with the explicit goal of informing both qualitative and quantitative evaluation criteria of different architectures against the open source science data system objectives. It was held virtually on Oct 19-20, 2021. Workshop #2 will focus on understanding the state-of-the-art in mission data processing systems and open science, as well as seeking community input on data system architectures. It is planned to be virtual in February/March 2022. More information on the state-of-the-art will be solicited by a broader community via invitation and an open RFI. After Workshop #2, the SAWG will conduct a system architecture trade study evaluating different architectures against the evaluation criteria. Over the 4-month period during this study, the SAWG will communicate with stakeholders from Workshop #1 to inform assessment of architectures for meeting different criteria. Workshop #3 will present candidate architectures and make a recommendation with an assessment against the criteria. This is planned for August 2022 in a virtual format. NASA Headquarters will then decide a path forward.



	2021							2022																		
	Oct		Nov		Dec			Jan		Feb		Mar		Apr		May		Jun		Jul			Aug		Sep	
Milestones	10-08	10-22	11-05	11-19	12-03	12-17	12-31	01-14	01-28	02-11	02-25	03-11	03-25	04-08	04-22	05-06	05-20	06-03	06-17	07-01	07-15	07-29	08-12	08-26	09-09	09-23
Workshop #1 (virtual)		🟩																								
Report Preparation #1			🟦	🟦																						
Prep for Study					🟨	🟨	🟨	🟨	🟨	🟨																
Workshop #2 (in-person KSC)											🟩															
Report Preparation #2												🟦	🟦													
Conduct Study														🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧					
Final Review (in-person GSFC)																						🟥				
Final Report Preparation #3																						🟦	🟦			
Workshop #3 (virtual)																								🟩		

Figure 1. A gantt chart of the project timeline.

Workshop #1 Summary Findings

For Workshop #1, the focus was on the input from various NASA Program Officers and ESO Missions regarding requirements, constraints, recommendations, and opportunities for science data processing. The two day workshop had Day 1 focused on programmatic presentations from the managers of all the NASA Earth Science Programs and Day 2 consisted of flight project presentations from the listed ESO projects who are in pre-phase A. After each group of presentations, the SAWG was able to ask questions directly to the presenters and each day ended with an open discussion for the totality of the daily presentations. In compliance with Open Science, the workshop was open to anyone, but the steering committee did request attendees to register for metric purposes. The workshop was also recorded and the recordings, as well as the presentations are hyperlinked and made available on the Study Website. The below figures show the makeup of the registered attendees as self identified during the registration process.

144 participants attending

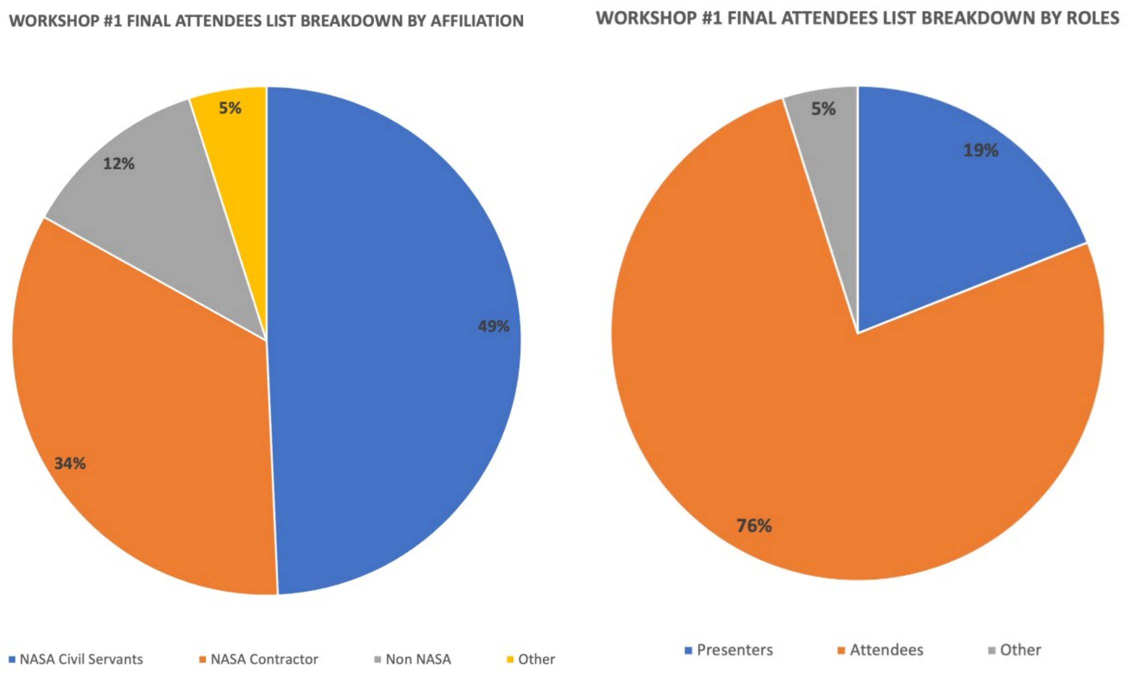


Figure 2. Workshop participants by affiliation and role.

Stakeholder Priorities

NASA Programs

There are several programs under NASA Earth Science Division (ESD), including: Flight, Research and Analysis (R&A), Applied Sciences Program (ASP), and the Earth Science Technology Office (ESTO). Each program supports NASA Earth Science Division objectives, but has different missions. Here we outline the general priorities across ESD, and then each subsequent program. General guidance across ESD involved setting a high-level agenda to prioritize open science. The Transform to OPen Science (TOPS) initiative targets a cultural shift for capacity building, partner engagement, and incentives to help accelerate scientific discovery through open source science (OSS). This includes development of supporting technologies and infrastructure to enable OSS with the intent to foster broad participation by NASA data users earlier in the mission scientific development process, starting in Phase B. By engaging through well-established technologies, supported by well defined processes and policies that are adopted across missions and across centers that consider emerging cybersecurity policies, NASA aims to engage a broader community.

To achieve OSS, NASA ESD is communicating their intent for OSS in a consistent and direct way to all relevant stakeholders with the idea that early communication can enable standardized access and trust. OSS can build trust through credibility in the scientific process that makes science transparent, accessible, inclusive, and reproducible. This is essential as NASA holds information [publications, data, & software] as a public asset to increase knowledge and serve the public good.

Specific to the Flight ("Project") Program, there are three driving priorities. First, any OSS data system must be realistic about how we can implement it within existing data systems design constraints and other flight requirements such as NPR 7120.5, NPR 7123.1, cyber policies, WBS, etc. Second, any data system must be compliant with existing concepts and investments (see glossary). Third, such a mission data system must enforce the role of Data Active Archive Centers (DAACs) as "Science Enabling Centers" and their services for data and metadata stewardship, information management, open source software support, cross-mission science and modeling, and user support.

Specific to the R&A Program, there are four driving priorities. First, any OSS data system must accommodate the breadth of activities including data from surface-based measurement networks, airborne instruments and platforms, scientific computing, global modeling, and calibration/validation infrastructure. Second, the R&A stakeholder community for which such a data system includes NASA centers, universities, federal laboratories, other government agencies, international partners, private sector entities, and non-profit institutions. Most of this community is supported through competed individual investigator awards, but there may also be some directed funding, especially for enabling activities at NASA centers. Third, R&A extends the open data policy by increasingly supporting publications in open journals as such R&A views OSS as collaboration for not just sharing code and intermediate products, but also as sharing knowledge. Fourth, OSS enables community-based, intra-team collaboration for: data analytics as well as software and tools using open source programming languages and infrastructure.

Specific to the Applied Science Program, there are two driving priorities. These priorities are motivated by their aim to expand connections with businesses, foundations, and nonprofit organizations, and international governments, while continuing to build upon our robust partnerships with government agencies. This program aims to enable people and organizations to apply insights from Earth science to benefit the economy, health, quality of life, and environment. The first priority is to recognize that work is done in partnership with user organizations, external to NASA and an OSS data system needs to enable co-development approaches with decision-maker, analyst, and scientist teams. Second, a data system must focus on enabling access to users of variable skill; e.g., teams include remote sensing scientists and managers on the ground collaborating on topics of emerging and urgent need. The purpose must be to improve the capabilities of individuals and institutions to access and apply NASA Earth Science data and information. This includes: use and integration of multi-mission data across disciplines, the ability to work with mission data and proxy data, and access to low-latency data. Meeting these priorities would amplify applications development, scaling, and access by decision-makers and the supporting organizations.

Specific to the Earth Science Technology Office (ESTO), there are four driving priorities. First, an OSS data system would include both on-premise (High-End Computing - HEC/High Performance Computing- HPC) and cloud capabilities. Second, an OSS data system would facilitate documentation and registration of algorithms, software, tools, and data so that adhere to FAIR data principles (Wilkinson et al., 2016). Third, an OSS would consider the mechanisms, processes, and policy that facilitate how technologies developed under ESTO are infused. There are two types of technologies: New Observing Systems (NOS) and Analytic Center Frameworks (ACF). Both of these are developed under the Advanced Information Systems Technology (AIST) program that identifies, develops, and supports adoption of software and information systems, as well as novel computer science technologies expected to be needed by the Earth Science division in the 5-10 year timeframe. AIST facilitates infusion and adoption and many technologies from different projects developing technologies that can be federated to work more generally across disciplines. Specific to an OSS data system, the ACF projects develop open source software platforms to facilitate adoption and evolution that facilitates OSS and accelerates generation of information products at scale. Specifically, ACF projects enable science users to share workflows, tools, and data tailored to needs of the ESO science community for visualization, ease of use, and shareable results. They focus on scalable parallelization through cloud, High-Performance Compute (HPC) processing and hybrid cloud architectures as well as consider model acceleration via surrogates. AIST projects demonstrate emerging technologies for analysis across instruments and scales through data fusion analytics, cloud-based ACFs and federation across projects and science data processing systems.

ESO Programs

Given that the first objective of the OSS data system architecture study is to “meet the ESO mission science data processing objectives”, each of the four ESO project program offices provided priorities specific to their ESO. The ESO project program offices are based out of NASA headquarters and are responsible for building a cohesive plan for that projects’ relevance to each program within NASA ESD. These program offices manage budgets across NASA centers involved in developing the mission as well as strategic partnerships such as partnerships with other federal agencies and international agencies.

The NASA-Indian Space Research Organization (ISRO) Synthetic Aperture Radar (NISAR) and Surface Deformation and Change (SDC) projects will collect data over all land and all ice on every orbit, including sea ice and coastal regions. NISAR supports research and applications that span nearly every Focus Area and Application Science Program Area and take advantage of dense time series. NISAR's priority for an ESO OSS data system is to lower the bar for data usability for all NISAR data and products in three ways. First, OSS can help develop the tools that grow communities (especially non-technical) through the creation and sharing of new tools and capabilities that target the diverse reach of the community including users that do not know that they would benefit from SAR/InSAR data products. Second, key to lowering the bar for data usability is the creation of L3+ products and cloud computing that brings the users to the staggering volume of NISAR data, about 140 PB (petabytes) in 3 years, a greater than 6x increase in the entire Earth Science Data and Information Systems (ESDIS) 2017 holdings. NISAR is already demonstrating how this could be done through the support for L3 NISAR product Algorithm Theoretical Basis Documents (ATBDs) that can be run for anywhere in the world by anyone, since project will only produce the products over calibration/validation (cal/val) sites. Third, NISAR data must be in a format and project that can be easily combined with other mission data in analysis-ready data (ARD) formats to facilitate data fusion and analysis like modeling and analytics.

The Surface Biology and Geology (SBG) project will collect data over all land and coasts. SBG supports research and applications that span nearly every Focus Area and Application Science Program Area. The priorities for the SBG program are to maintain an open dialogue between the missions and OSS initiative to ensure that good intent doesn't inadvertently preclude important partnering mechanisms necessary for the SBG observing system constellation that includes multiple international partners. SBG intends to begin a two-way dialogue with historically excluded communities across all fronts – in the OSS initiative; at the mission level in the diversity of the teams, outreach, and implementation, and in our program management at HQ. The purpose is to enable the breadth of SBG-facilitated science by including a diverse user base with variable experience, which relies on OSS.

The Mass Change (MC) project will continue and improve the gravity measurement observational record from GRACE (2002 to 2017) and GRACE-FO (launched in 2018). MC recognizes the role of level 3 (L3) data products significantly support Earth System science both through direct climate record (20+ years of data) and through data assimilation. The priority for MC for an OSS data system is recognition that not all algorithms can be open source, for example the L1-L2 product algorithms are proprietary, while the L3+ products are more flexible and able to be shared through OSS.

The Atmosphere Observing System (AOS) - formerly the Aerosol, Cloud, Convection and Precipitation (ACCP) - project explores the fundamental questions of how interconnections between aerosols, clouds, and precipitation impact public health, weather and climate, addressing real-world challenges to benefit society. AOS aims to understand the processing of water and aerosol through the atmosphere and develop the societal applications enabled from this understanding. AOS measurements are taken at multiple times a day from an inclined orbit segment and at lower cadence from a polar orbit segment, and includes cal/val data as part of a science-driven suborbital program, long-term regional networks, supersites, and targeted campaigns. AOS primarily focuses on resolving process understanding and secondarily on extending existing climate data records. AOS is one observing system that has two space segments with suborbital measurements from aircraft and the potential for additional contributions from other space agencies and their observing systems: JAXA, CSA, CNES. The priority for the AOS program for an OSS data systems is that it needs to accommodate a diverse ground data system from ~8 sources and CSA Inuvik participation as well as external partners for science and applications including calibration and validation (cal/val)l partners and the broader research community (both NASA-funded PIs and other).

ESO Science Teams

Given that the third objective of the OSS data system architecture study is to “support Earth system science and applications”, each of the four ESO project science teams provided priorities specific to their ESO. The ESO project science teams represent leaders in the scientific community who are experts in the measurements specific to each ESO and who work closely with the variety of scientific disciplines and applications communities for which that ESO serves. These science teams are responsible for providing input and guidance on mission formulation to enable use of the data by the broader communities for which they represent.

The NISAR/ SDC project science team identified four priorities for an OSS data system that would support Earth System science and applications enabled by NISAR. The baseline scope of NISAR data processing will support cross-disciplinarity domains of surface deformation and change, ecosystems, cryosphere, and also applications community needs such as urgent response. With this in mind, the first priority is that all ROSES PIs and Agency partners need access to all software on github/lab and mirrored cal/val data with computational resources provided through NASA cloud assets co-located with the data, e.g. Alaska Satellite Facility’s OpenSARLab/HyP3. Second, algorithms for L3 products can be contributed by the broader community without vetting, but the algorithms adopted by the science team would be vetted and included in the gitlab for use by the community. Third, project scientists feel that instant open access with appropriate quality metadata is most conducive to scientific progress. Fourth, it is necessary to enable semantic searches for cross-mission data combinations in a data and algorithm catalog. This would include quality and resource utilization metrics that are standardized and databased, to be used in machine learning assisted workflow assistance and guidance as well as standardization to visualize algorithms (workflow graphs), data products, and science results. Finally, cal/Val and science advances require overlap with previous mission (GFO) and other ancillary data sources: Satellite Laser Ranging measurements, GNSS data from Precise Orbit Determination of satellites in LEO (e.g. SWARM), Ocean Bottom Pressure Recorders, GNSS data (Earth surface deformation), extrapolation of existing mass change data record (GRACE, GFO), known variations in regions of low variability, satellite altimetry and ARGO over the ocean and ice sheets, estimates of terrestrial water storage reconstruction products, and model-enhanced products of of mass variations (hydrology and ocean). As such, NISAR algorithm development will require access to these data.

The SBG project science team identified three priorities for an OSS data system that would support Earth System science and applications enabled by SBG. First, SBG is a constellation of two satellites: 1) VSWIR imaging spectrometer and 2) a TIR radiometer with VNIR camera that are collaborating with other missions for data harmonization: ESA LSTM TIR radiometer, ESA CHIME VSWIR spectrometer, and CNES/ISRO TRISHNA TIR radiometer. This collaboration is essential to meet science and applications needs for reduced revisit and observing events. This requires a need for open data sharing and product harmonization with international partners CHIME (VSWIR), LSTM and TRISHNA (TIR) and a need for international collaboration on cal/val of terrestrial and aquatic networks for vicarious cal/val activities on six continents. Second, the OSS system must enable and support use of interoperable ESO and POR data to create data harmonization. Finally, the SBG Mission would benefit from a cross-ESO common science and applications platform where open source science activities may be enabled with frictionless cross-mission data fusion for analytics, algorithm testing and data production.

The MC project science team identified three priorities for an OSS data system that would support Earth System science and applications enabled by MC. First, the project, instrument and flight operations need immediate access to L0 data in order to provide science team access to L1 data within 14 days and L2 within 40 days (typical) while providing quick looks within 1 day, but latency is not considered a driving requirement. Second, the project expects to produce L1 to L3 products with value for L3-L4 products produced by the user community including project international partner agencies. Third, the MC science team relies on 20 years of maturing algorithm development to produce the products they produced.

The AOS project science team identified six priorities for an OSS data system that would support Earth System science and applications enabled by AOS. First, the AOS project-funded science team that includes international pattern agencies (JAXA, CNES, and CSA) will produce L0 to L4 “at launch” standard products. These include derived products that use both a single instrument and multiple instruments. Second, an OSS data system should enable experimental products post-launch to be developed by ROSES or open science. Third, an OSS data system would enable L3 data products to be used in the atmospheric data assimilation system. Fourth, the OSS data system must enable data product latencies on the order of 3-6 hours for L1-L3 spectrometer data; 2 hours for L1-L3 radar, polarimeter, and LIDAR data; and 1 hour for L1b radiance data from the radiometer. Fifth, to enable early engagement of modeling/scientific/applications user communities, AOS plans to have a pre-launch Analytic Collaborative Environment with simulated data from Airborne "Satellite simulator" instruments. This environment would enable sharing of documentation, provision of data production environment and algorithm code/libraries. Sixth, the value of AOS data to applications (decision support) is expected to come from cross-mission synergies.

ESO Mission Data Processing Systems

Given that the second objective of the OSS data systech architecture study is to “enable data system efficiencies”, each of the four ESO project science teams provided priorities specific to their ESO mission data processing systems. The ESO mission data processing systems (MDPS) are the set of algorithms, software, compute infrastructure, operational procedures, and documentation to automatically process raw instrument data through to science quality data products. This includes the software tools that support the development of the processing algorithms and validation and analysis of the processed data. The MDPS receives data from the Ground Data System (GDS) that acquires data from on-board the flight platform (satellite or aircraft). After processing data in the MDPS, data is delivered to the NASA Data Active Archive Center (DAAC) for long-term archive and user access. The MDPS is traditionally accessed by the Project teams including mission operators, software developers, and science teams.

The NISAR/SDC MDPS had seven priorities. First, all L0-L2 software needs to come with complete documentation and tests with a full continuous integration (CI) process. Community developments can be evaluated for compliance and adopted as needed. The InSAR Scientific Computing Environment (ISCE) platform development over the past 10 years is an example of how this can work. Second, based on the volume of data that NISAR will be generating, efficient algorithms will be necessary/required to prevent a data processing backlog. This includes the use of multi-core and GPU-enabled computing resources. This requires a need to understand the source data for each ESO mission and when to transform it to a product (data format, projections, L3, L4+) that can be combined and integrated with other data types; for example, L1 single look complex (SLC) as compared to radiometric-terrain correction (RTC) data products, which can be treated as another “band” in a data stack. This would make the data accessible by a broader user community as analysis-ready data (ARD). Third, data formats should also be compatible with sister missions (e.g. all SAR missions for a SAR constellation). NISAR data products will be stored in the ASF DAAC and the NISAR MDPS enables an easy, effective switch from MDPS production to DAAC archive. Given the high data volumes (100+PB), this requires collocation of the GDS, MDPS, and DAAC in the same AWS region. Fourth, NISAR compute needs fluctuate greatly based on reprocessing needs and has a low latency need (urgent response) for some products. NISAR MDPS data system efficiency includes the ability to optimize for lower-cost processing pipelines (e.g. bulk [re]processing) while providing scalability for latency sensitive processing pipelines (e.g. forward "keep-up" processing and urgent response). Fifth, all software and data need to be available in the open (as allowed by organizational policies). NISAR is using AWS, custom c++ code and python, and standard open source tools like gdal, numpy, and eigen in containerized conda environment instances. Sixth, there is a need for “on-demand” higher-level product generation, which requires easy access to cal/val data not just mission data including all metadata. Seventh, the NASA and ISRO teams use different systems and there is no plan for a common production platform, but there will be a data sharing portal.

The SBG MDPS had five priorities. First, SBG has a high number of potential products (>200 in 10 product suites), which requires an innovative open science MDPS solution as most users do not have the skills to work with L2+ data products that are GIS-ready (“application ready”). To enable mission selection products, the MDPS must support an algorithm sandbox that is open and enables “on demand” L3+ products by users thereby increasing the user base and accelerating transition SBG data to providing societal benefits. Second, the SBG MDPS is moving away from proprietary and legacy code and using open code repositories and tools. All software and data will be available in the open as allowed by organizational policies. Third, the SBG MDPS needs to have the ability to optimize for lower-cost processing pipelines (e.g. bulk [re]processing) while providing scalability for latency sensitive processing pipelines (e.g. forward "keep-up" processing and urgent response). Low latency data products are required for many

SBG applications. Third, the SBG MDPS will have multiple components at different NASA centers and international partners (JPL, AMES and ASI). Fourth, the SBG MDPS must be able to support high data volumes (100+PB), which will require deploying it on AWS so that it is co-located with the DAAC to enable efficiency between processing and archive. Fifth, the SBG MDPS must enable data harmonization with international agency partners and the program of record. This will require collocation of data from many sources, rapid ingest of partner data, cross-calibration and data fusion to common time/space grids.

The MC MDPS had six priorities. First, an MC MDPS would allow flexibility to infuse new technologies and enable optimal observable development and future observing system evolution (future double satellite pair vs single pair) while building on the GRACE/GRACE-FO MDPS with components spanning NASA centers (JPL and GSFC) as well as German partner GFZ. Importantly, L1+ code is not shared since each MPDS has their own unique software packages. This diversity in software/approaches is of great value to: test parameterizations and implementations, contribute to validation and verification. Second, an MC MDPS would facilitate interfaces with other agencies and international mission partners and services (e.g. operational services). Third, the MC MDPS processing needs require supercomputing resources with large amounts of intermediate data storage (e.g. current local compute cluster is 500 TB - mostly full), but final data product volumes are small (<1PB expected). The current infrastructure uses an on-premise compute fleet; while the project explored AWS, it was cost prohibitive. High-end Computing is a possibility. Fourth, an MC MDPS needs access to proprietary code for L1-L2 products, but openness for L3+ products is welcome and could benefit cal/val and support Earth system science. While open science for an MDPS would allow unrestricted access of data and tools to broad users, existing MC L1-L2 workflows have limited portability and openness. Portability is easy to address with software engineering practices, but not without huge investment because the software suites used to process L2 data have been developed over decades and consist of hundreds of thousands of lines of code. This would be a big investment given that there is no clear need identified in over 20 years of GRACE/GRACE-FO data to make L1-L2 code open. There are also licensing issues with software; for example, GRACE-FO Caltech considered code developed at JPL proprietary. Fifth, an efficient MC MDPS would provide the research community and other users validated standard mission products that avoid duplication of product development and storage, while recognizing that 77% of users want L3+ products as demonstrated from PO.DAAC downloads. Sixth, there are some low latency needs (urgent response) for some products.

The AOS MDPS had five priorities. First, the AOS MDPS is expected to produce 5.1 PB/year (~15 PB total) excluding multi-instruments products. Second, the MDPS will be managed by GSFC and located in AWS, co-located with DAAC. Third, the AOS MDPS will leverage algorithms and software developed for legacy missions and all software and data will be available in the open as allowed by organizational policies. Third, software will be developed with configuration management and automated continuous integration, delivery, and deployment. Algorithm software will be decomposed into reusable parts. Fourth, the AOS MDPS must support low latency needs (urgent response) for some products. Fifth, to enable early development of simulation data and processing, the AOS MDPS will include an Analytic Collaborative Environment.

Stakeholder Considerations

Innovation is a change in the current business operations to new business operations, which requires systematically evaluating the existing system as it pertains to Policy, Economics, Sociocultural Factors, and Technologies/Tools (PEST) (Aguilar, 1967; Stavros, 2021). As such, we provide a synthesis of key considerations identified by workshop participants with respect to these four factors. The findings here represent overarching concerns relevant to all stakeholders.

Policies

Policies to consider include cybersecurity, intellectual property (IP), and International Traffic in Arms Regulation (ITAR), as well as SMD policies on Data Management for Groundbreaking Science, Scientific Information, and Open Source Science.

Cybersecurity policies are a big hurdle to enabling open science. Specifically, thinking about how to enable cross-organizational and public access to organizational resources like data and processing. Also, there are considerations for quality control requirements by NASA as they relate to the Data Quality act.

Intellectual property is nuanced and varies by partner. The OSS data system needs to be sensitive to the fact that partners (especially private sector but also international) may have different approaches to sharing information (e.g., proprietary nature). An OSS data system would enable flexibility for how to handle IP.

ITAR regulation controls the manufacture, sale, and distribution of defense and space-related articles and services as defined in the United States Munitions List (USML). The goal of the legislation is to control access to specific types of technology and their associated data. Overall, the government is attempting to prevent the disclosure or transfer of sensitive information to a foreign national for national security. An OSS data system needs to consider implications on national security.

The NASA Science Mission Directorate (SMD) Strategy for Data Management and Computing for Groundbreaking Science 2019-2024 aims to: 1) lead an innovative and sustainable program supporting NASA's unique science missions with academic, international and commercial partners to enable groundbreaking discoveries with open science; and 2) continually evolve systems to ensure they are usable and support the latest analysis techniques while protecting scientific integrity. In compliance with this policy, an OSS data system architecture would: a) develop and implement a consistent open data and software policy tailored for SMD; b) upgrade capabilities at existing archives to support machine readable data access using open formats and data services; c) establish standardized approaches for all new missions and sponsored research that encourage the adoption of advanced techniques; and d) partner with academic, commercial, governmental and international organizations.

The SMD Policy Document SPD-41, Scientific Information policy for the Science Mission Directorate is a consolidation of existing policies applicable to SMD. These policies are based on our understanding of existing NASA and Federal guidance, and are relevant for all current and future NASA awards. This applies to all SMD funded activities related to producing scientific information, but the policy excludes restricted information. Most important to an OSS data system architecture is that: a) SMD-funded data shall be made publicly available without fee or restriction of use; b) SMD-funded software should be released as open-source software; c) all SMD-funded activities shall have data management plans (DMP) describing the management and release of data to facilitate the implementation of these information policies. The DMP should include a description of the software to be used and how it will be managed. Policy updates (applied by 2022) will include requirements to adhere to Findable, Accessible, Interoperable, and Reproducible (FAIR) data principles (Wilkinson et al., 2016) and requirements for persistent identifiers (e.g., DOIs) attached to grants.

The SMD Open Source Science Policy specifies policy within the context of SMD Policy SPD-41 with relevant data system mandates that: 1) all mission data, metadata, software, databases, publications, and documentation be available on a full, free, open, and unrestricted basis starting in Phase B with no period of exclusive access or conditions for use; 2) software be developed openly in a publicly accessible, version-controlled platform using a permissive software license allowing for community use and contributions; 3) NASA and [Partner] software, documentation and data be properly marked, cited, and/or attributed with metrics to measure and acknowledge open-source science contributions will be developed; and 4) all research data shall become publicly available no later than the publication of the peer-reviewed article that describes it including data and software (released as open source software) required to derive the findings communicated in figures, maps, and tables.

Economics

The economics of an OSS data system that must be considered include cost-constraints, cost allocations, value proposition, funding grant life cycles, and previous investments in NASA cyberinfrastructure. First, an OSS data system is cost-constrained (but not cost-capped) by ESO budgets. Second, there needs to be a mechanism for cost allocations for data access, storage and processing as used by the programs, projects, and end users. Third, the value of working with the OSS data system must exceed the costs. While NISAR is all-in on cloud computing; they are just scratching the surface on tools to best exploit cloud systems and the community is still convinced it is cheaper to buy their own systems and download the data, the largest problems are not being tackled because of data availability now, and data access/cost in the future. Third, the funding grant life cycle limits longevity of open source software. There is a need to anticipate expected demands by “code users” and plans to provide continued support from originators that accounts for algorithm/code updates, funding status, and/or personnel changes. This may include being sure that teams providing widely shared codes are funded to provide needed support, including documentation. Alternatively, there needs to be clear communication to the community that they should not expect that support. Finally, NASA has invested in HEC/HPC that has access to specialized computing systems focusing on parallel computing tasks that are not cost-effective in commodity based computing infrastructure.

Sociocultural Factors

To implement an OSS, there are sociocultural factors that must be considered including the existing business model for the science community, the current science reward system, teaching new practices to experienced researchers, and developing a consistent OSS standard across centers.

The science community core business is based on the reward for being the first to discover, invent, or develop, with little appreciation for spending extra resources making things available to others, documenting, and answering questions, which would slow down the core business. This has been a problem for HEC program that found algorithms, software and systems were not documented to facilitate sharing. This also requires overcoming PI concerns about ideas/research getting scooped. It’s necessary to treat OSS as intellectual property and normalize software development as equal intellectual merit as algorithm development.

This leads to a need to change the culture and create incentives to participate in OSS. This will be essential as PIs and NASA centers compete for funding resources. Incentives include making the process as easy as possible. With NASA’s move of new mission data in AWS, a successful OSS program would have a Cloud entry point that is very compelling for a researcher to make the transition from local to the Cloud (Why is it in their best interest? What do they gain? Some incentives of such a system would provide tools and capabilities that they may not have access to (further incentives) and inclusion in the development of new OSS tools/algorithms provided there were mechanisms to promote their work and get credit for past contributions.

This leads to the challenge and need to teach experienced researchers new practices. For example, we need to socialize user communities and provide trainings that ease them into the cloud and HPC to enable them to fully exploit the efficiencies and potential.

Finally, creating a cultural shift will require developing consistent OSS guidelines across the different centers (e.g. academia, other NASA agencies, etc.) and disciplines to improve community understanding and compliance.

Technologies/Tools

Technological considerations include use of existing investments in infrastructure, data tools as much as cyberinfrastructure, and mechanisms for community contributions.

OSS must consider existing platforms, developments, policies, and lessons learned into a new initiative supported by new technologies. This involves leveraging existing investments in HEC Program that sustains the most advanced cyberinfrastructure, which provides computational resources to NASA funded projects including large scale computing tasks. HEC has a reference architecture, used when soliciting projects, for enabling open source software in support of open source science.

The OSS must not only consider cyberinfrastructure, but also the open-source software tools to enable data search, discovery, access and use. Researchers need to be able to find the tools/solutions they seek and have quality metrics for assessing tool utility. For the ESOs, there is a level of complexity of data processing at L1-L2, that requires users to have specialized expertise. This is why projects would typically not promote products from individual users. In particular, MC's past experience shows that too many flavors of data products (or tools) can be detrimental to expanded uptake by the community by causing confusion with new users. Less can be more. This would involve some kind of code quality metrics, as well as algorithm quality. Algorithm quality is typically verified in a controlled environment to test viability and performance and assessed with cal/val data to test algorithm viability. This may involve allowing users access to the same production tools that the Science Teams use for data product quality assessment and validation. However, there is a lot of value in community contributions. For example, NASA user communities (e.g. application users) may use specialized formats that are not a part of standard NASA practices. Community-contributed tools can help these users gain access to NASA data in a less complex format where they can handle using simpler capabilities (e.g. spreadsheets).

Not only must the OSS enable use of open-source software tools, it must also provide a mechanism for users to contribute their inputs/feedback into the OSS system. For example, how do we bring in new capabilities and contributions to development? How do we bring in community contributed algorithms and incorporate these inputs back into the MDPS? How does the OSS allow researchers who are not a part of the mission to contribute? How do we handle algorithms contributed by international partners in the context of ITAR, IP and cybersecurity policies? One consideration in answering these questions is how to handle interterm versions of software, documentation and data products? These should be developed openly, possibly resulting in different release "channels" (e.g. Beta channel, Alpha channel, stable, etc.).

Stakeholder Constraints

Constraints on any new data system architecture must minimally provide what currently exists. As such, we provide a SWOT analysis as a synthesis of Strengths, Weaknesses, Opportunities for improvement, and Threats to implementing improvements as presented by workshop participants. Responses represent consensus across stakeholders.

Strengths

Strengths of existing data systems are that they include activities to make NASA data accessible, support of large data and compute infrastructure, and have known costs and schedules. NASA has already invested in activities for making NASA data accessible including Cloud-ready data, Openscapes, the SMD Data Catalog Search, Standards for NASA Data, and AI-Enabled data. They have also invested and actively support large data and computing infrastructure designed for the most demanding computational workloads. These systems enable data processing and analytics. This computing infrastructure routinely refreshes data and is maintained. There is often idle compute at HPC facilities, due to the nature of large simulations jobs, and this can be used to complete data processing jobs that are often embarrassing parallel. The advantage to leveraging this existing infrastructure is that many existing MDPS use it, so the implementation is known and thus is ready to meet schedules with known costs.

Weakness

Weaknesses of existing data systems include limitations on community access, limitations on use of existing infrastructure, challenges with interoperability across compute infrastructure, lack of expertise to work with NASA data, and lack of recognition and merit for support OSS software and tooling.

At present, cybersecurity policies limit access to open science platforms by non-NASA users (e.g. NSF, NOAA), in part because of NASA rules on who can access NASA systems. This leads to challenges with sharing large data sets. Additionally, it leads to inequitable access depending on who is authorized to access NASA assets (e.g. cloud and compute resources). For example, the NISAR United States Science team and Cal/Val partners use a project-funded production system. Presently the system operates within the JPL firewall, which limits access to science team and partners that cannot be on-boarded to operate within the JPL firewall because of barriers such as: 1) security background checks and 2) difficult processes to sign up for and maintain two-factor authentication, tokens, passwords.

While NASA has invested a lot in cyberinfrastructure through HEC, HPC mostly supports modeling and simulation (MODSIM). NASA HEC/HPC are currently not optimized for the high-disk I/O characteristics that some ESO missions (NISAR and SBG) would need. However, there may be an architecture that supports both cloud and HPC. This would consider developing optimized data transfer mechanisms between these computing resources.

One challenge with working across HPC and cloud is that it has historically required software development to port algorithms and make them interoperable across computing platforms. Seamless interoperability across computing resources doesn’t exist today. While containerization and virtualized environments help with portability, in other environments versioning of different operating systems and build systems for open source software leads to fragility. Furthermore, representation of a scalable compute in the cloud and HPC systems are historically different approaches. Modification of algorithms and workflows are historically needed to more maximally leverage the efficiencies of HPC and cloud.

Another weakness of existing infrastructure for enabling broader access is because of the extremely high skill required for data processing, which results in a small community of experts globally. For example, in the case of MC, “the satellite is the instrument”, so L1/L2 processing requires an expert user with sophisticated knowledge of the entire spacecraft system. Furthermore, production of data products requires expertise in software engineering and code optimization as it can be a large effort to migrate scientific code to production code. For example, MC requires the ability to reprocess 20+ years of data in an efficient manner, repeatedly, for necessary experiments.

Finally, a major weakness of existing infrastructure is a lack of intellectual merit given equally to algorithm developers and software developers and there is not equal recognition for converting science code into open source software. This is in part because of the reward systems in place (See considerations under Sociocultural Factors).

Opportunities

Data system opportunities identified include capacity building, early adopters, standardization of the value of OSS, use of existing infrastructure, and developing open source software and data tools.

There is an urgent need to build capacity among the broader research community both for OSS and for use of any NASA infrastructure built. For this end, we can leverage professional societies (e.g., AGU, AMS) and partnerships (ESIP) for training, policy developments, and exchange of new ideas as well as the NASA Applied Science capacity building programs (ARSET, DEVELOP, and SERVIR). In developing capacity, there is a need to target engagement with those communities that are less well positioned to use/access data/technology.

Important to building capacity is engaging the community early and gaining champions in the research community who can help pave the way to widespread adoption. The term “early adopter” is used widely in the technology adoption literature to mean: “a person who starts using a product or technology as soon as it becomes available”. At present, NASA thinks of “early adopters” as those who use mission data early and may be involved in algorithm development during mission data system development. These early adopters are seen to provide valuable feedback on community requirements for higher level products and improve development timelines. However, there are opportunities to expand Early Adopters to include adoption of the cloud. Given NASA’s commitment to move future mission data archive into AWS, there is a need for all missions to have Open Science/Cloud Computing Early Adopter Programs. Early career and graduate students can act as Early Adopters and uptake of OSS (e.g., through hackathons). Additionally, early adopters can help OSS and to develop collaborative “end user” data tools. Finally, when encouraging early adoption, there is a need to target engagement with communities that are less well positioned to use/access data/technology.

OSS offers an opportunity to increase usability of NASA data for the broader community by developing tools that support data access and curation as well as pre-processing (or data “munging”). This can include tools that help collocate related geophysical variables temporally and spatially, visualize and display basic statistics, contextualize data with other geopolitical variables, identify anomalies, etc. To enable community contributions to tool development there are two key needs. First, datasets and cod indeed quality and resource utilization metrics to enable machine-learning assisted workflow management to efficiently enable exploration of massive data sets and the breadth of available tools. Second, there is a need to provide science/tech partnering opportunities to enable open-source software development and support.

There is also an opportunity to leverage past investments by NASA in High-End Computing (HEC). HEC HPC could become part of the next generation MDPS specifically for data reprocessing and modeling. Many ESOs need an MDPS to have the ability to optimize for processing pipelines that are not latency sensitive (e.g. bulk [re]processing), as well as optimize for latency sensitive processing pipelines (e.g. forward "keep-up" processing and urgent response). With this in mind, lower latency processing can be done in the cloud, while non-latency sensitive workloads may be able to be done on existing NASA HEC/HPC infrastructure. Implications of cloud egress costs need to be accounted.

Finally, NASA has the opportunity to set an example for our national/international partners for OSS policy and encourage all to adopt it. To do this, we would need to make open sharing a criteria for a competitive proposal. We need to consider developing metrics such as an o-index (similar to h-index) that qualifies the impact of PI's contributions to OSS. This would make open science a scorable metric that could be used to evaluate competitive OSS proposals in the selection process. In making OSS a part of funding solicitations, there is a need to provide funding/training to those communities that are less well positioned to use/access data/technology.

Threats

Threats to providing an ESO Open Source Science Data system include a disconnect between OSS policy and implementation within the existing business model and cybersecurity policies; the potential to introduce inequity in terms of access, inclusion, and work effort; cost-optimization in the cloud and across diverse geographic regions; and research community hesitancy on cloud compute adoption.

At present, there is a disconnect between HQ OSS policy and the realities of the existing science business model and cybersecurity policy. Specific to the science business model, there are currently a lack of financial incentives to participate in OSS, especially as it relates to competition for funding. There is a need to ensure scientists' contributions to OSS (e.g., sharing of algorithms, codes, etc.) can be tracked to give appropriate recognition by agencies and the research community, which involves having it be factored into selection criteria for awards and promotions. With respect to cybersecurity, there are limitations to open up systems, software, and data to the public. Great progress has been made in NASA to create open source software, but hurdles still exists to make platforms accessible by users for open algorithm development and science analysis. Cyber policies often flow down from other government institutions, there might be limits on what can be done with respect to openness.

A potential threat to an ESO data system that enables OSS is the potential to introduce inequity in terms of access, inclusion, and work effort. Specifically thinking about access and constraints to accessing NASA systems introduced by cybersecurity policies, there is a need to consider who is getting access to ensure that not only the well-supported and well-connected are positioned to use products and services (e.g., data tools) developed by others. With respect to inclusion and work effort, open source software development and approval process is time-consuming. This can put strain on marginalized groups that are unable to maintain both discipline excellence and technological savvy. Finally, those who do provide OSS resources like data tools may do a lot of work, but get little credit as others come in later use the tool and do not properly credit the creator. Care needs to be taken to mindfully consider how an ESO OSS data system might have unintended consequences introducing inequity and how it can use metrics and tracking to give proper credit to OSS contributors.

One threat to an ESO OSS data system that serves multiple ESO's with teams that are geographically scattered and catering to a global research community is the challenge of cost-optimizing data storage and use in the cloud. It can be challenging to enable discovery and low-cost access of data in physically distinct cloud regions. Specifically, egress costs and time to move data out of a common data lake by NASA (e.g. in AWS us-west-2 Oregon region) limits efficient processing elements outside of the common region. There may need to be a centrally-controlled, distributed computing solution where data reduction operations are conducted near large-volume data, and lower-volume data is then transferred between regions. In contrast to data systems using the cloud, existing MDPS not in the cloud already provide predictable costs and schedules.

Finally, a potential threat to an ESO OSS data system is the difficulty of technology adoption and education by the broader science community to work in a "new" or different system. There is a need to provide demonstrable advantages to any new "solutions" to drive and support adoption. There is generally a lack of training in OSS and a resistance to usage due to unfamiliarity with an application, service, etc. A technology plan needs to be developed and considered from the beginning.

Synthesis of Common Themes by Data System Objective

Table 1 describes a synthesis of common themes by Data System Study objective.

Table 1. Summary of common themes for identifying criteria for evaluating architectures based on study objectives (left column).

Objectives	Common Themes
ESO mission science processing objectives [“Mission Dev”]	1. Facilitating the production and incorporation of L3+ products produced by the community is necessary. 2. The production of low latency products must be supported. Where low latency is defined as time between acquisition and data access by the user; specifically this does not relate to urgent response as time between event and acquisition. 3. The infrastructure will either be AWS or project owned on-prem compute. 4. Interfacing with external, potentially international, MDPS's is necessary. 5. Missions expect to produce mission-products through Level 2. 6. Data system must account for architecture maturity and feasibility to meet ESO timelines. 7. Data system is cost-constrained by ESO mission budget capacity.
Enable Data System Efficiencies	1. Co-locating with the DAAC is important in reducing data duplication and data transfer time/costs 2. Must be flexible to efficiently (cost, bandwidth, processing capability) support large and small data volumes 3. A scalable infrastructure to accommodate variable compute needs over time is crucial to reducing costs 4. Processing data in a format that is widely accessible 5. Have ability to optimize scalability for latency sensitive processing pipelines (e.g. forward "keep-up" processing and urgent response) different from standard workflow.
Support Earth System Science	1. Ability to share data across ESO missions is beneficial for enhancing common science goals. 2. Access to data from other missions (other than ESO) is crucial. 3. Ability to share algorithms across ESO missions. 4. Must support multidisciplinary research by providing core services to facilitate cross-disciplinary data integrations. 5. Higher level products (L3+) are necessary for studying the Earth System 6. A common ESO data system architecture may amplify project's Earth system science objectives
Promote Open Science principles	1. Users need access to an on-demand science/applications analysis platform to create [custom] L3+ data products. Many needs also include on-demand of any products, not just L3+ 2. Users ability to use the platform without sharing software and code publicly. 3. Ability to manage and incorporate community contributions, feedback, & new capabilities 4. Ability to track metrics for users open source science contributions to encourage adoption by the community 5. Ability to implement standardized open source science guidelines (e.g., metadata standards, provenance, etc.) 6. Cost accounting is needed to determine who pays for access, storage, and processing in the cloud

Path Forward

After synthesizing the needs of different Stakeholders in creating a data system architecture to support the ESO missions that considers data system efficiencies while supporting Earth system science and applications as well as open science principles, the SAWG developed a set of evaluation criteria (Table 2). These evaluation criteria are considered design constraints by which to evaluate different architectures. This term is used in place of “requirements”, which are often traced for data systems from higher-level mission requirements; hence the avoidance of prescribing them for all ESO and future missions. For each “common theme” numbered in Table 1, there are one to two evaluation criteria listed in Table 2.

The next steps for this study are to synthesize and consolidate these eval criteria into a concise list of design constraints. Deconstruct a data system architecture into bounding conditions and key components. The second workshop and Request for Information (RFI) will solicit input from the broader community of different data systems (and their components) and their potential for integration into different architecture based on their feasibility, maturity, and dependencies. Based on this input, we will amass different candidate architectures that can be assessed against the different evaluation criteria. Several candidate architectures will be presented to NASA Headquarters for programmatic review and potential selection as a path forward. Essential to integration and uptake of any architecture is continued and sustained engagement with each ESO as their schedules evolve.

Table 2. Based on the common themes from stakeholders (Table 1), we define evaluation criteria as design constraints for evaluating data system architectures. They can be defined and reported as quantifiable and qualitative metrics.

Objectives	Data System Evaluation Criteria as they correspond with Common Themes from Table 1
ESO mission science processing objectives [“Mission Dev”]	1. The Data System shall provide an accessible platform for the community to develop L3+ products that adhere to NASA metadata and provenance standards 2. The Data System shall support on-demand product generation as soon as data is available from the ground data system. 3. The Data System shall be portable to support deployment on-prem, in-cloud, multi-cloud, and hybrid infrastructure. 4. Two: 1) The Data system external interfaces shall go through Authentication and Authorization. 2) The Data system external interfaces shall use standardized access protocols. 5. Two: 1) The data system shall have the ability for forward-stream and bulk [re-]processing. 2) The Data System shall be compliant with DAAC archive retrieval. 6. The Data system shall be demonstrated (TRL6+) by the earliest ESO mission launch. 7. The Data system is cost-constrained by ESO mission budget capacity.
Enable Data System Efficiencies	1. The Data System shall support a data lake. 2. The Data System shall be flexible to efficiently (cost, bandwidth, processing capability) support large and small data volumes. 3. The Data system shall accommodate variable compute needs over time is crucial to reducing costs. 4. The Data System shall support services to create standard data formats (ESDIS Standards) 5. The Data System shall keep up with forward-stream processing demand.
Support Earth System Science	1. The Data System shall enable ESO data sharing before data archive. 2. The Data System shall enable data access from non-ESO missions. 3. The Data System shall enable users to share algorithms. 4. The Data System shall enable development and sharing of data tools (e.g., software, code libraries, etc). 5. The Data System shall enable on-demand processing. 6. The Data System shall meet cross-ESO mission science goals.
Promote Open Science principles	1. Two: 1) The Data System shall provide an analysis platform. 2) The Analysis Platform shall be accessible by NASA and Non-NASA users while supporting algorithm sharing and on-demand batch processing. 2. The Analysis Platform shall enable users to control user-generated resources: private and public code repositories, containers, binaries, data, etc. 3. The Analysis Platform shall enable public access. 4. The Analysis Platform shall enable user authentication and authorization for provenance of contributions. 5. The Analysis Platform shall automate standardized open source science guidelines (e.g., metadata standards, provenance, etc.). 6. The Data System shall allocate cost accounting by user activity. 7. The Data System shall be compliant with cybersecurity policies (e.g., authenticate and authorize, user management, etc.). NOTE: Intellectual Property compliance is covered by other evaluation criteria.

References

Aguilar, F. J. (1967). *Scanning the Business Environment*. Macmillan Publishers Limited.

Margetta, R. (2021, May 24). New NASA Earth System Observatory to Help Address Climate Change [Text]. Retrieved June 7, 2021, from http://www.nasa.gov/press-release/new-nasa-earth-system-observatory-to-help-address-mitigate-climate-change

NRC. (2019). *Thriving on Our Changing Planet: A Decadal Strategy for Earth Observation from Space: An Overview for Decision Makers and the Public* (p. 25437). Washington, D.C.: National Academies Press. https://doi.org/10.17226/25437

Ramachandran, R., Bugbee, K., & Murphy, K. (2021). From Open Data to Open Science. *Earth and Space Science*, *8*(5), e2020EA001562. https://doi.org/10.1029/2020EA001562

Stavros, E. N. (2021). Wicked Problems need WKID Innovation: Innovation as a Process to Develop a Disruptive Technology Product. *Qeios*. https://doi.org/10.32388/HQBGX6

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Glossary

Analysis-Ready Data - are satellite data that have been processed to a minimum set of requirements and organized into a form that allows immediate analysis with a minimum of additional user effort and interoperability both through time and with other datasets.

Application - use of NASA data for decision support (policy, resources, etc).

Application-Ready Data: GIS-ready data

Accessible - Data, tools, software, documentation, publications follow FAIR Data Principles.

Cloud - Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. (NIST SP 800-145, 2011)

Capability Need - functionalities of the system.

Common ESO data system - a standardize MDPS that services multiple ESO missions.

Data Active Archive Centers (DAACs) - are NASA data archives that serve different research communities but share common services to standardize NASA data management and archive through the NASA Earth Science Data and Information System (ESDIS).

Data Lake - The concept of data-proximate processing where the data is stored and co-located with the processing.

Data Product Level - All definitions are assumed to be consistent with the NASA Data Processing Levels: https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy/data-levels

Earth System Observatory (ESO) - a constellation of satellites will be launched by NASA in the 2020s to observe the Earth System as designated by the National Academies Decadal Survey (NRC, 2019) and classified as “designated observables”.

Evaluation Criteria - are design constraints by which to evaluate different architectures. This term is used in place of “requirements”, which are often traced for data systems from higher-level mission requirements; hence the avoidance of prescribing them for all ESO and future missions.

FAIR Data Principles - Data should be Findable, Accessible, Interoperable, and Reproducible by machines (Wilkinson et al., 2016).

Ground Data System - The system responsible for receiving telemetry data from the observatory and providing it to the MDPS, which does the instrument specific processing.

Hybrid Cloud - Infrastructure that is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability. (NIST SP 800-145, 2011)

Inclusive - The process and participants welcome participation by and collaboration with diverse people and organizations.

Latency - defined as time between acquisition and data access by the users.

Mission Data Processing System (MDPS) - The set of algorithms, software, compute infrastructure, operational procedures, and documentation to automatically process raw instrument data through to science quality data products. This includes the software tools that support the development of the processing algorithms and validation and analysis of the processed data.

On-prem Computing - Computing infrastructure that physically resides within an enterprise owned data center, server room, etc. On-prem may be referred to as “in-house”. Usually, an organization is fully responsible for procuring, deploying and managing on-prem computing.

Open Science - “a collaborative culture enabled by technology that empowers the open sharing of data, information, and knowledge within the scientific community and the wider public to accelerate scientific research and understanding” (Ramachandran et al., 2021).

Open Source Science - builds on concepts from Open Source Software revolution that expanded participation in developing code and applies it to the scientific process to accelerate discovery through open science from project initiation through implementation.

Open Source Software - The Open Source Initiative (OSI) defines Software to be Open Source if distributed under a license with a set of criteria: 1) license shall not restrict any party from selling or giving away the software, i.e. free redistribution, 2) source code is included with any program or set of programs, 3) license allows for derived works, 4) integrity of author’s source code, 5) licence must not discriminate against a groups or persons, 6) license must not discrimination against fields of endeavor, 7) any rights must apply to all whom a program or source is redistributed to, 8) rights attached to the program must not depend on the program's being part of a particular software distribution, 9) license must not place restrictions on other software that is distributed along with, and 10) no provision of the license may be predicated on any individual technology or style of interface. (https://opensource.org/osd)

Permissive software - software that can be copied, modified, redistributed, etc.

Reproducible - The scientific process and results can be reproduced by members of the community.

System Architecture Working Group (SAWG) - a team of system engineers, data system architects, software engineers, and ESO mission representatives tasked with conducting the ESO open source science data system architecture study. The SAWG is composed of science data system experts who represent the diversity of the data system community and are connected to the end-user science community and the ESO missions.

Steering Committee - the leadership team for the ESO open source science data system architecture study responsible for providing programmatic insights and steering the SAWG to conduct a programmatically relevant study.

Transparency - Both the scientific process and results are visible, accessible and understandable.

Acronyms

ACCP	Aerosol, Cloud, Convection, and Precipitation
ACF	Analytic Center Frameworks
AGU	American Geophysical Union
AI	Artificial Intelligence
AIST	Advanced Information Systems Technology
AMS	American Meteorological Society
AOS	Atmosphere Observing System
ARCO	Analysis-Ready Cloud-Optimized data
ARD	Analysis-Ready Data
ARSET	Applied Remote Sensing Training
ASF	Alaska Satellite Facility
ASI	Agenzia Spaziale Italiana
ASP	Applied Sciences Program
ATBD	Algorithm Theoretical Basis Documents
AWS	Amazon Web Services
cal/val	Calibration and validation
CHIME	Canadian Hydrogen Intensity Mapping Experiment
CNES	Centre National d'Etudes Spatiales
CSA	Canadian Space Agency
DAAC	Distributed Active Archive Center
DEVELOP	Digital Earth Virtual Environment and Learning Outreach Program
DMP	Data Management Plan
DOI	Digital Object Identifier
ESA	European Space Agency
ESD	Earth Science Division

ESDIS	Earth Science Data and Information System
ESIP	Earth Science Information Partners
ESTO	Earth Science Technology Office
ESO	Earth System Observatory
FAIR	Findable, Accessible, Inter-operable, Reproducible
GDS	Ground Data System
GFO	Gravity Recovery and Climate Experiment Follow-On
GIS	Geographic Information System
GNSS	Global Navigation Satellite System
GPU	Graphics Processing Unit
GRACE	Gravity Recovery and Climate Experiment
GRACE-FO	Gravity Recovery and Climate Experiment Follow-On
GSFC	Goddard Space Flight Center
GFZ	Geoforschungszentrum
HEC	High-end Computing
HPC	High Performance Computing
HQ	Headquarters
InSAR	Interferometric Synthetic Aperture Radar
IP	Internet Protocol
JAXA	Japan Aerospace Exploration Agency
JPL	Jet Propulsion Laboratory
ISRO	Indian Space Research Organisation
ISCE	InSAR Scientific Computing Environment
ITAR	International Traffic in Arms Regulations
L#	Data Product Level # as defined by https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy/data-levels

LEO	Low Earth Orbit
LIDAR	Light Detection and Ranging
LSTM	Long Short Term Memory
MC	Mass Change
MCP	Microsoft Cloud Platform
MDPS	Mission Data Processing System
MODSIM	Modeling and Simulation
NASA	National Aeronautics and Space Administration
NISAR	NASA-ISRO Synthetic Aperture Radar (SAR)
NIST	National Institute of Standards and Technology
NOAA	National Oceanic and Atmospheric Administration
NOS	New Observing Systems
NPR	NASA Procedural Requirements
NSF	National Science Foundation
OSI	Open Source Initiative
OSS	Open Source Science
PB	Petabytes
PEST	Policy, Economics, Sociocultural Factors, an Technologies/Tools
PI	Principal Investigator
PO.DAAC	Physical Oceanography DAAC
POR	Program of Record
R&A	Research and Analysis
RFI	Request For Information
ROSES	Research Opportunities in Space and Earth Sciences
RTC	Radiometric-Terrain Correction (SAR Data product)
SAR	Synthetic Aperture Radar
SAWG	System Architecture Working Group

SBG	Surface Biology and Geology
SDC	Surface Deformation and Change
SDS	Science Data System
SLC	Single Look Complex (SAR data product)
SMD	Science Mission Directorate
ST	Science Teams
SWOT Analysis	Strength, Weakness, Opportunity, Threat
SWOT	Surface Water and Ocean Topography
TB	Terabyte
TCO	Total Cost of Ownership
TIR	Thermal Infrared
TOPS	Transform to OPen Science
TRISHNA	Thermal infraRed Imaging Satellite for High-resolution Natural resource Assessment
USML	United States Munition List
VNIR	Visible and Near-Infrared
VSWIR	Visible to Short-Wave Infrared
V&V	Validation and Verification
WBS	Work Breakdown Structure